From the world to word order

From the world to word order: Deriving biases in noun phrase order from statistical properties of the world

Jennifer Culbertson, Marieke Schouwstra, Simon Kirby

Language, Ahead of Print, vol. 96, no. 3, (Article)

Published by Linguistic Society of America
DOI: *https://doi.org/10.1353/lan.0.0245*

FROM THE WORLD TO WORD ORDER: DERIVING BIASES IN NOUN PHRASE ORDER FROM STATISTICAL PROPERTIES OF THE WORLD

Jennifer Culbertson
Centre for Language Evolution, University of Edinburgh

Marieke Schouwstra
Centre for Language Evolution, University of Edinburgh

Simon Kirby
Centre for Language Evolution, University of Edinburgh

The world's languages exhibit striking diversity. At the same time, recurring linguistic patterns suggest the possibility that this diversity is shaped by features of human cognition. One well-studied example is word order in complex noun phrases (like *these two red vases*). While many orders of these elements are possible, a subset appear to be preferred. It has been argued that this ordering reflects a single underlying representation of noun phrase structure, from which preferred orders are straightforwardly derived (e.g. Cinque 2005). Building on previous experimental evidence using artificial language learning (Culbertson & Adger 2014), we show that these preferred orders arise not only in existing languages, but also in improvised sequences of gestures produced by English speakers. We then use corpus data from a wide range of languages to argue that the hypothesized underlying structure of the noun phrase might be learnable from statistical features relating objects and their properties conceptually. Using an information-theoretic measure of strength of association, we find that adjectival properties (e.g. *red*) are on average more closely related to the objects they modify (e.g. *wine*) than numerosities are (e.g. *two*), which are in turn more closely related to the objects they modify than demonstratives are (e.g. *this*). It is exactly those orders which transparently reflect this—by placing adjectives closest to the noun, and demonstratives farthest away—that are more common across languages and preferred in our silent gesture experiments. These results suggest that our experience with objects in the world, combined with a preference for transparent mappings from conceptual structure to linear order, can explain constraints on noun phrase order.*

Keywords: word order, typology, silent gesture, corpora, information theory

1. Introduction. One of the oldest debates in linguistics concerns whether the languages of the world share a set of core invariant properties reflecting universal features of human cognition. At the center of this debate is a tension between the diversity we see when we look across languages and the similarities that crop up when they are analyzed under a certain lens. This tension, between linguistic diversity on the one hand and universal organizing principles on the other, is on full display in one of the simplest linguistic structures we use: the noun phrase. Given just a noun (e.g. *vases*) and three common categories of words that modify it—a demonstrative (e.g. *these*), a numeral (e.g. *two*), and an adjective (e.g. *blue*)—there are already twenty-four possible ways of ordering the words to make a phrase, almost all of which are found in some language. For example, the English order is *these two blue vases*; in Thai, it would be the equivalent of *vases blue two these*; in Vietnamese, it would be *these two vases blue*; in Basque, it would be *two vases blue these*; and so on. Yet there remains a small subset of orders that no language appears to use systematically. For example, we currently know of no language that systematically uses the equivalent of *blue two these vases* or *blue these vases two*.

Linguists have argued that these missing patterns offer evidence of universal organizing principles underlying how noun phrases are built (Cinque 2005, Steddy & Samek-Lodovici 2011, Abels & Neeleman 2012, Dryer 2018, Steedman 2018). As careful

* We would like to thank Roger Levy and the referees for their comments on previous versions of this work. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 Research and Innovation Programme (grant agreement No. 757643).

analyses of noun phrase order exist for only a small sample of the world's languages (around 700 in Dryer 2018), any one pattern could be absent by chance (Piantadosi & Gibson 2014). Here, we focus not on which patterns are currently attested in the world's languages, but instead on the frequency differences among the twenty-four possible orders. The dramatically skewed distribution is shown in Figure 1a.[1]

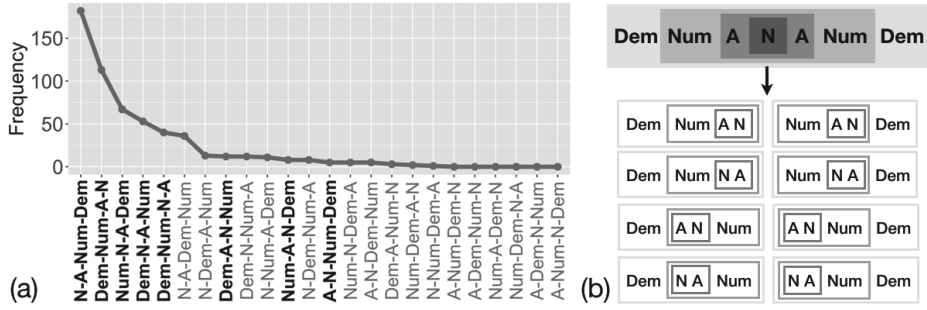

FIGURE 1. (a) Estimated frequency of each of the twenty-four possible orders (based on counts of languages sampled in Dryer 2018; N: noun, A: adjective, Num: numeral, Dem: demonstrative), with homomorphic patterns highlighted in bold black, showing a clear preference for homomorphism across languages. (b) Schematic representation of subunits or constituents in the noun phrase and the resulting eight homomorphic orders, which preserve this underlying structure.

2. HOMOMORPHISM AND NOUN PHRASE WORD ORDER. What sort of organizing principles might explain why some noun phrase orders are so much more common than others? All current accounts start from the idea that adjectives, numerals, and demonstratives are not created equal. Rather, they differ in how they combine with each other. To illustrate this, take a complex word like *speakers*, composed of a lexical root *speak* and two morphemes *-er* and *-s*. The meaning of the word reflects how these two morphemes combine with the root; *speak* combines with *-er* first, creating a noun, *speaker*. This larger unit is then pluralized by combining with *-s*. The order of semantic composition is here preserved in the linear order—the morpheme that combines its meaning with the noun root first is closer to the root. This same idea can be applied to see how elements in the noun phrase (here multiple words) combine to form a coherent meaning. The adjective forms a unit with the noun first (i.e. *vase* is modified by the property *blue*). The resulting unit then combines with the numeral (i.e. the numerosity of the blue vases is specified), and finally that unit combines with the demonstrative (e.g. the group of blue vases is located in space relative to the speaker). This composition order is typically assumed to be reflected in the syntax (Adger 2003, Alexiadou et al. 2007), creating an underlying hierarchical structure in which each subunit forms a syntactic constituent.[2] Just as in the case of morpheme order, the linear order of words in a noun

[1] Different typological samples and sampling techniques reported in the literature result in slightly different frequency estimations, including some discrepancies as to which orders are unattested (Cinque 2005, Cysouw 2010, Dryer 2018). Importantly, the shape of the distribution remains the same regardless of the method. For example, raw frequencies from Dryer 2018 are displayed in Fig. 1a, but a measure of frequency that aims to correct for genetic and geographic distance still shows a handful of common orders, homomorphic orders among them, with a long tail of infrequent orders (Dryer 2018).

[2] Basic constituency tests also show that in a phrase like *these blue vases*, *blue vases* is a constituent (it can be replaced by *ones*). This order of composition is supported by formal semantic accounts of different modifier types. For example, most adjectives are treated as predicates that combine with the noun (e.g. Partee

phrase can in principle reflect this underlying structure, or not. An order that does will have the adjective placed closer to the noun, and the demonstrative farthest away. Following Martin et al. (2020), we refer to these as HOMOMORPHIC orders.[3] There are eight such orders, shown in Figure 1b, and they make up the bulk of the most frequently attested orders in Fig. 1a.

The notion of homomorphism—a transparent mapping between underlying structure (i.e. the compositional units described above) and linear order—thus describes a kind of hidden similarity between languages that on the surface appear to be different. This is exactly the kind of universal organizing principle posited by many linguists, but it is worth unpacking what this might mean. If the explanandum is the frequency differences among noun phrase orders, then two potential organizing principles must be involved. First, there is a universal preference for transparent mappings between underlying structure and linear order. A universal preference is not a hard-and-fast constraint, but rather one that is, by hypothesis, present in all humans but violable in their languages (e.g. as in Culbertson et al. 2013). After all, the majority of languages are homomorphic, but non-homomorphic languages can arise and are learnable. In addition, there is reason to believe that transparent mappings are preferred across cognition, reflecting a domain-general preference for simplicity in learning (Chater & Vitányi 2003, Culbertson & Kirby 2016).

The second piece of the puzzle is the underlying structure itself, in particular, the compositional units described above. Some linguists have argued that constraints on noun phrase order provide potential evidence for innate knowledge (Cinque 2005, Abels & Neeleman 2012, Steedman 2018). While the most obviously language-specific constraints proposed in these theories are designed to rule out specific non-homomorphic orders (rather than to explain the high frequency of homomorphic ones), underlying these theories is the universality of the hierarchy. Where does this structure come from? One possibility is that the categories Adjective, Numeral, Demonstrative, and Noun are innately known (or expected) by language learners, who tacitly know how they combine semantically, and thus come to the task of language acquisition already equipped with an underlying syntactic structure based on this (Adger 2003). In other words, from the moment children map words in their language onto these categories, they will expect Adjectives to combine with the Noun before Numerals, and Demonstratives to combine last.

1987), while numerals are widely analyzed as functions from nominal predicates to countable units (e.g. Partee 1988, Heim & Kratzer 1998) and demonstratives as functions mapping nominal predicates to individuals (e.g. Elbourne 2008). This also aligns with functionally oriented work on the noun phrase, which argues that nouns and adjectives appear closer together syntactically than nouns and numerals do because they are closer semantically (Hurford 1987, Rijkhoff 1990, 2004). All previous accounts of noun phrase word order cited here therefore assume this underlying structure (in terms of either syntax or semantics, or both).

[3] The term ISOMORPHIC is used in Culbertson & Adger 2014, but as Martin et al. (2020) point out, it is more accurate to call these orders homomorphic, reserving isomorphic for the two most frequent orders—N-Adj-Num-Dem and Dem-Num-Adj-N. These are the only two orders from which it is possible to fully recover the underlying structure. In Dem-Num-N-Adj, by contrast, the surface order does not contradict the structure illustrated by Fig. 1b, but it is not possible to recover the relations between, for example, Dem and Adj. Therefore such orders are homomorphic but not isomorphic. The especially high frequency of isomorphic orders could then be explained by an independent preference for word-order harmony. Indeed, both Culbertson et al. (2012) and Dryer (2018) argue that there is a preference for consistent placement of modifiers before or after the noun. The two orders N-Adj-Num-Dem and Dem-Num-Adj-N are the most common because they are homomorphic and have a consistent order of modifiers relative to the noun. Of the non-homomorphic orders that are attested, many are harmonic (e.g. N-A-Dem-Num, N-Dem-A-Num, N-Dem-Num-A).

Here we explore these two hypothesized universals—a preference for homomorphism, and a universal underlying structure for the noun phrase (reflecting semantic composition and/or syntactic constituency). First, we show that when English speakers improvise a system of gestural communication, their gesture orders are systematically homomorphic. This supplements existing experimental evidence for a homomorphism bias in humans (Culbertson & Adger 2014, Martin et al. 2020) and supports the claim that this bias is at play in explaining noun phrase order in established languages. Then, we use an information-theoretic measure of strength of association to argue that the universal structure that shapes noun phrase order may in principle be learnable from observing the world, rather than reflecting innate knowledge. Specifically, we show that objects are more closely associated with their properties than with their numerosities; objects and their numerosities are in turn more closely associated than objects and their location and/or relation to the speaker. These nested conceptual representations (which are not linguistic in nature), combined with the linguistic categories Noun, Adjective, Numeral, and Demonstrative, form the basis of the hierarchy from which noun phrase linear order is derived.[4] The skewed distribution of orders across languages may thus come from a pressure to be homomorphic combined with a universal hierarchical structure derived (in part) from properties of the world around us.

3. Experiment 1. While previous accounts of noun phrase word order have implicitly assumed that homomorphic orders are a kind of default (e.g. Abels & Neeleman 2012), recent work has sought to provide direct behavioral evidence for a homomorphism bias using laboratory experiments (Culbertson & Adger 2014, Martin et al. 2019). For example, Culbertson and Adger (2014) trained English speakers on a pseudo-artificial language: participants saw English phrases with a prenominal modifier (e.g. *blue vases*, *two cows*, *these shoes*) and heard a translation into the new language, where modifiers were postnominal (e.g. *vases blue*, *darts two*, *shoes these*). Crucially, the phrases they were trained on only ever had a single modifier (either an adjective or a numeral or a demonstrative); multiple-modifier phrases were held out, so no evidence was given about the relative order of modifiers in the new language. At test, participants were shown these held-out, multiple-modifier phrases in English (e.g. *these two vases*, *two blue vases*, *these blue vases*) and asked to guess how they should be translated in the new language. One option was consistent with the surface order of modifiers in English, but not homomorphic (e.g. *vases these blue*). The other featured the reverse order of modifiers from English, but was homomorphic (e.g. *vases blue these*). Participants consistently chose the homomorphic option over the alternative, despite the latter being more probable given their experience with strings of modifiers in English. This is in line with a homomorphism bias operating on a universal underlying structure. However, the results could also reflect transfer at a more abstract level; participants may have learned from English that surface order should be homomorphic, and may have transferred this to postnominal modifiers in the experiment.

Stronger evidence would come from showing that speakers of a non-homomorphic language still show a bias favoring homomorphic order. However, this is challenging for two reasons. First, very few well-documented languages are in fact non-homomorphic.

[4] Here we remain agnostic about whether and how children acquire the linguistic categories themselves. Furthermore, the centrality of the Noun in the hierarchy is likely something to be explained. For example, it may result from an object or shape bias, again either learned or innate (Landau et al. 1988, Kucker et al. 2019). Here we restrict our attention to the learnability, from nonlinguistic properties of the world, of the particular hierarchical nesting of conceptual representations.

Second, widespread bilingualism means that even if we were to test those speakers, they are still very likely to have experience with a homomorphic language (whether English or otherwise). Instead, we attempt to bypass the effects of prior linguistic knowledge to the degree possible by using the silent gesture paradigm. In silent gesture experiments, participants with no knowledge of a sign language must improvise a way to convey information using only their hands and no speech. This method has been popular in exploring biases underlying basic word order, showing that when participants use gestures to describe simple events (e.g. *Alex kicked the ball*), they often bypass the dominant order of their native language (Goldin-Meadow et al. 2008, Futrell et al. 2015) and take the semantic or conceptual properties of the information to be conveyed into account (Gibson et al. 2013, Hall et al. 2013, Schouwstra & de Swart 2014, Schouwstra et al. 2016). Here we use this method to investigate biases in noun phrase order. We expect that, as in experiments on basic word order, gestures will not simply recapitulate English order. Rather, if a bias for transparent mapping between underlying structure and linear order is at play even when participants are improvising in a modality distinct from their previous language experience, we expect their gesture order to be homomorphic.

3.1. METHOD.

PARTICIPANTS. Participants were twenty native English speakers. Data from four participants were excluded due to failure to produce gestures containing information for more than one modifier; therefore data from sixteen participants were used for analysis. None had previous knowledge of any sign language.[5]

MATERIALS. The stimulus set consisted of images of squares or triangles. They appeared in groups of four or five and were either striped or spotted, as in Figure 2a. Location relative to the gesturer was represented by two iPads that displayed the images, one of which was directly in front of the participant, and the other about an arm's length away. The eight different images, presented on two different iPads, together formed sixteen total stimulus items.

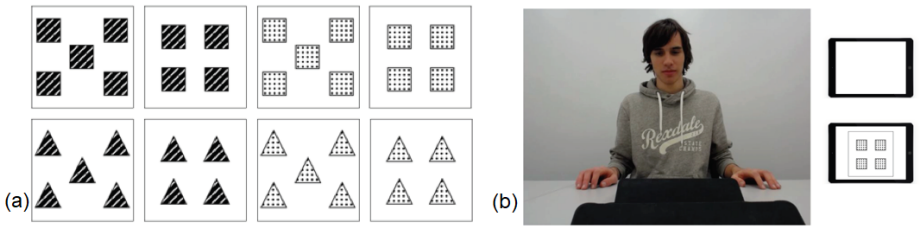

FIGURE 2. (a) Stimulus set for experiment 1. (b) Set-up of the experiment, with participant seated in front of two iPads displaying images (e.g. as shown alongside).

PROCEDURE. Participants were seated at a table across from the experimenter, with the two iPads in front of them (see Figure 2b). They were filmed using a Logitech camera connected to a Macbook Air, which controlled stimulus presentation over a networked server. Before starting, participants were shown the set of eight images they would have to gesture as printed pictures. They were told that they should describe each image using their hands, without any speech, so that someone watching the recorded video would be able to work out what they were seeing. The sixteen total items (eight

[5] All experiments reported here were approved by the PPLS Ethics Committee at the University of Edinburgh. All participants gave consent prior to beginning. This included permission to share videos/images from their sessions for research purposes.

images in two locations) were presented twice to participants in two randomized blocks. There was a brief break after the first block of sixteen trials.

Coding. Example gesture clips are shown in Figure 3. Gestures were coded by identifying which part of the gesture corresponded to information associated with each conceptual piece of the image: the object corresponded to the N(oun), the texture/size to the A(djective), the numerosity to Num(eral), and the location to Dem(onstrative). In some cases, additional information other than that corresponding to the N, A, Num, or Dem was conveyed (e.g. about the spatial arrangement of the objects in an image). This was ignored for the purposes of coding order. If Dem, A, Num, or N were omitted, or gestures included simultaneous information (i.e. if the object shape was gestured while simultaneously indicating numerosity), no order was coded for the relevant element(s).[6] Gesture strings were then coded as homomorphic or not, based on whether they were consistent with one of the eight orders shown in Fig. 1b. For example, Dem-N-A-Num was homomorphic, as was N-A-Num (with omitted Dem); Dem-N-Num-A, by contrast, was non-homomorphic, as was N-Num-A. Gestures were coded once by the first author, and a second time by an independent coder. Agreement on gesture order was 82%, and 92% on gesture homomorphism. All instances of disagreement were resolved by discussion with a third coder.

Figure 3. Clips from two participants illustrating experimental set-up with proximal and distal iPads and example gestures. (Top: stimulus *four spotted squares* on proximal iPad, order Dem-Num-N-A; bottom: stimulus *four striped triangles* on distal iPad, order Dem-N-A-Num.)

3.2. Results. As predicted, gesture strings were overwhelmingly homomorphic; indeed, only 2% were non-homomorphic (11% were ambiguous, e.g. due to gesture combinations or repetitions). To confirm this, we ran a logistic mixed-effects model predicting a binary outcome variable, homomorphic or not, using only an intercept term and including participant as a random effect.[7] This intercept was positive and significant ($\beta = 4.71 \pm 0.91$, $p < 0.001$), confirming a strong preference for homomorphic gesture strings. The frequency of gesture orders and how often a given pattern was used in the majority of a participant's gestures (i.e. as the participant's default order) are shown in Figure 4a and 4b, respectively. All of the latter are homomorphic. The most com-

[6] In a small number of cases (seven total), simultaneous gestures meant we could not determine whether the order was homomorphic (e.g. in a gesture string with Dem followed by a simultaneous gesture incorporating N, A, and Num, the relative order of A and Num is not clear).

[7] All regression models reported here were run using the lme4 package (Bates 2010) in R (R Core Team 2017). All data reported in this article, as well as analysis scripts, are available at https://osf.io/nurwp/.

monly used patterns were Dem-Num-N-A and Dem-N-A-Num, neither of which is the basic noun phrase word order used in English (Dem-Num-A-N). This raises the question, as in other studies employing the silent gesture paradigm, of what these gesture strings correspond to linguistically. While they could correspond to more complex noun phrases (e.g. akin to the English phrase *these two squares with spots*), they could also correspond more closely to sequences of sentences (e.g. *On that iPad, there are two squares. The squares have spots*). Alternatively, it could be that there is not a clear equivalent to speech, at least in the mind of the gesturer. While we return to this point below, these data suggest that, regardless of the linguistic status of these gestures, the information conveyed is organized temporally in a way that is homomorphic to the hypothesized underlying structure.[8]

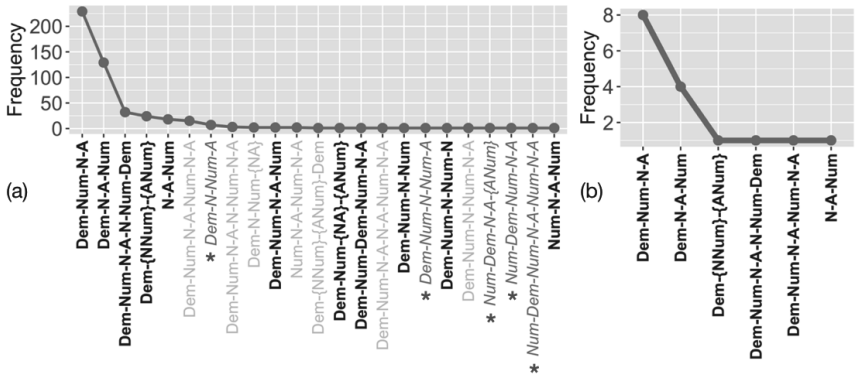

Figure 4. (a) Raw frequency of each order produced by participants. Bold black patterns are homomorphic; gray italic patterns preceded by stars are non-homomorphic; light gray patterns are ambiguous; elements in curly brackets were gestured simultaneously. (b) Patterns used in the majority of at least one participant's gestures (x-axis), and the number of participants using each (y-axis). The overwhelming majority of orders produced were homomorphic; none straightforwardly reflect the basic English order (Dem-Num-A-N).

4. LEARNING CONCEPTUAL STRUCTURE FROM THE WORLD. To summarize, typological evidence points to the high frequency of homomorphic orders relative to non-homomorphic alternatives. Artificial language learning experiments using spoken language suggest that English speakers assume homomorphic orders of postnominal modifiers (Culbertson & Adger 2014, Martin et al. 2019). This same population innovates homomorphic orders when improvising gesture sequences. All three sources of evidence point to the existence of a homomorphism bias, assuming an underlying structure in which the adjective combines with the noun first, then this constituent combines with the numeral, and then that larger constituent combines with the demonstrative.

We now turn to the origins of the underlying structure itself, which forms the basis of previous accounts of noun phrase order (Cinque 2005, Steddy & Samek-Lodovici 2011, Abels & Neeleman 2012, Dryer 2018, Steedman 2018). We propose that this structure is derived from universal conceptual representations that can be learned by observing the environment: objects in the world (expressed linguistically by the category Noun) are more closely related to their properties (expressed by Adjective) than to their nu-

[8] There is also a strong preference for gesturing the adjective information after the noun information. This accords with the typological counts of NP order, which show a prevalence of N-Adj types (Dryer 2018). A preference for postnominal adjectives has been discussed in the context of previous experimental work on noun phrase word-order learning (Culbertson et al. 2012, Culbertson & Newport 2015).

merosities (expressed by Numeral), which are in turn more closely related to objects than the objects' location and/or relation to the speaker is (expressed by Demonstrative). Intuitively, differences in strength of association among modifier types can be seen by considering common objects in the world. For example, wine is closely associated with its color (e.g. red or white); skyscrapers are closely associated with their height; Dalmatians are closely associated with their texture (e.g. spotted). It is trivial to come up with other such examples of objects that are closely associated with particular properties. By contrast, examples of objects that are closely associated with their numerosity are difficult to come by. Some things typically come in pairs, like shoes, or dozens, like eggs; however, most objects are not closely associated with a numerosity. Objects are even less likely to be associated with their location and/or relation to the speaker; in fact, these concepts are by their very nature changeable. These universal conceptual representations form the basis of the syntactic hierarchy, which by hypothesis specifies relations among discrete linguistic categories (an idea we return to below).

Strength of association can be formalized in information-theoretic terms as pointwise mutual information (PMI), given in 1. PMI tells us whether a given pair of elements cooccur more than would be expected from their base frequency rates.[9] If wine cooccurs with the property red more often than it would in a world in which objects and properties combined freely, then this pair will have high PMI. PMI for a pair of elements will be zero when the two elements are completely independent of one another, and negative when they cooccur less than would be expected by their base rates. Our prediction is that on average, objects and their properties have higher PMI than objects and their numerosities, which in turn have higher PMI than objects and their location and/or relation to the speaker.

(1) $$pmi = log \frac{p(x,y)}{p(x)p(y)}$$

To test our prediction, we estimated the PMI of objects and their properties, objects and their numerosity, and objects and their location/relation to speaker from dependency-parsed natural language corpora, on the assumption that the use of nouns, adjectives, numerals, and demonstratives in a corpus reflect the statistical properties of the world. We use corpora rather than, for example, image sets, because they provide a representative sample of the kinds of concepts that frequently appear (and are salient) in our environment. By contrast, large image sets are almost exclusively tagged with object names and do not include, for example, information about an object's relation to a speaker. Note, however, that using corpora means there will be some influence of linguistic categories, which are imposed on the world by our minds (see appendix §A1 for additional discussion). We calculate the base frequency rates of nouns and modifiers, and then count how often individual adjectives, numerals, and demonstratives modify individual nouns. The critical comparison is the average PMI values across (Noun, Modifier) pairs, for each modifier type. To ensure our results are not influenced by any particular language, we replicate this using corpora from twenty-four languages across a number of families, plus all English corpora in CHILDES (Sagae et al. 2007) (see appendix §A1 for details).

[9] Our account assumes that the relevant relations are between objects and their properties (or numerosities, or relation to the speaker). This builds in the centrality of the object (see n. 2). However, PMI is otherwise a symmetric measure: it therefore captures the intuition not only of a strong association between a Dalmatian and its spots, but also of the possibility that spots are very likely to bring to mind Dalmatians.

4.1. METHODS. We first extracted all dependencies involving a noun and an adjective, a noun and a numeral, or a noun and a demonstrative. We excluded all singleton pairs to prevent PMI values from being skewed by low-frequency items (Jurafsky & Martin 2019), and then estimated the probabilities of each pair, and each member of the pair. The probabilities were calculated using maximum likelihood estimation. The precise implementation of the PMI calculation was as in 2, where n is a noun, m is a modifier, and $t(n, m)$ is a noun and a modifier in a dependency of type t (either A, Num, or Dem). In other words, the probability of a given modifier is based on the set of modifiers of that type modifying a head noun, and the probability of the noun is based on the set of nouns that have that type of modifier.[10]

$$(2)\quad pmi = log\ \frac{p(n,m|t(n,m))}{p(n|t(n,m))p(m|t(n,m))}$$

Stepping through an example, say we have a corpus that results in the cooccurrence frequencies in Table 1a. The probabilities for the pairs (wine, red), (wine, spotted), (dog, red), and (dog, spotted) are given by their cooccurrence frequencies divided by the total counts of all (noun, adjective) pairs in the corpus (i.e. 1000/2011 = 0.497 for (wine, red) and (dog, spotted), 1/2011 = 0.0005 for (wine, spotted), and 10/2011 = 0.005 for (dog, red)). To obtain the PMI values for each pair, these numbers are divided by the individual probability of the noun in the pair (1001/2011 = 0.498 and 1010/2011 = 0.502, respectively), multiplied by the individual probability of the adjective in the pair (1010/2011 = 0.502 and 1001/2011 = 0.498, respectively), and we take the log of this number. The resulting PMI values are shown in Table 1b. The final step in our analysis is to compare the average of these pairwise PMI calculations for all noun-modifier pairs for each modifier type.

	wine	dog
red	1,000	10
spotted	1	1,000

a. Counts of cooccurrence frequencies.

	wine	dog
red	0.992	−5.665
spotted	−8.961	0.992

b. Resulting PMI values.

TABLE 1. Example of high- and low-PMI pairs from a made-up corpus with two nouns and two adjectives.

4.2. RESULTS AND DISCUSSION. Figure 5 shows average PMI values for each modifier type across all language corpora. As predicted, on average, (Noun, Adjective) pairs have the highest PMI, (Noun, Demonstrative) pairs have the lowest PMI, and (Noun, Numeral) pairs fall in between. A linear mixed-effects regression model predicting PMI from modifier type (a factor with three levels: Adjective, Numeral, Demonstrative) with language as a random effect confirms this (using Helmert contrast coding, with adjective as the default level: adjectives vs. numerals: $\beta = -2.44 \pm 0.02$, $p < 0.001$; mean of adjectives and numerals vs. demonstratives: $\beta = -0.16 \pm 0.01$, $p < 0.001$). Based on this information-theoretic measure of strength of association, properties (conveyed by Adjectives) are on average more closely associated with objects (conveyed by Nouns) than numerosities are (conveyed by Numerals), which are in turn more closely related with objects than location or status relative to the speaker is (conveyed by Demonstratives).

[10] Note that this is not the only way to implement the PMI calculation. For example, it is also possible to calculate probabilities over the set of all the phrases with a modifier modifying a head noun (i.e. across the whole set of noun phrases modified by A, Num, and Dem). The implementation reported here generally lowers PMI across the board relative to this alternative. However, both ways of computing PMI give the same results: adjective pairs have the highest PMI, and demonstrative pairs the lowest. For additional discussion of the relationship between PMI and entropy, see the appendix.

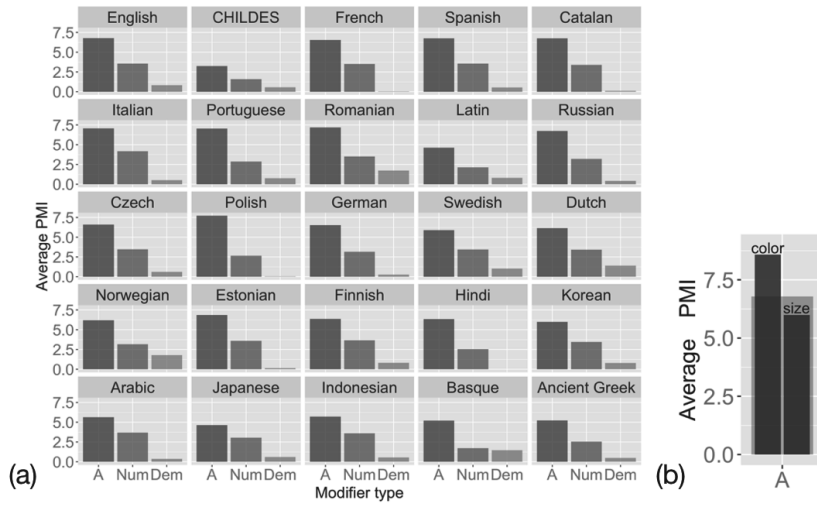

FIGURE 5. (a) Average PMI values across corpora of twenty-five languages confirming that, according to our measure, properties are more closely related to objects than numerosities are, and numerosities are more closely related to objects than location/discourse status is. (b) Distribution of PMI values for adjectives in the English corpora highlighting subset of color and texture vs. size adjectives (displayed as two bars overlaid on top of the average Adj PMI bar for English); by this measure, color/texture properties are more closely related to objects on average than size properties are.

Recall that we are using linguistic corpora here only as a convenient tool for getting at the statistical properties we assume to exist in the world, independently of language. How would these properties come to influence word order? Our hypothesis is that children track these differences in conceptual closeness while learning about the world, and use them to construct a representation of how objects and their properties, numerosities, and location relative to the speaker relate to each other. This nonlinguistic knowledge constitutes the conceptual basis for a natural asymmetry between the linguistic categories Adjective, Numeral, and Demonstrative, leading to the hierarchy in Fig. 1b. Importantly, this assumes that there is a level of representation—a linguistic hierarchy encoding syntactic constituency and/or semantic composition of Nouns, Adjectives, Numerals, and Demonstratives—mediating between conceptual representations of the world and word order. Given that there is independent evidence for the reality of such a level of representation, perhaps this is a reasonable assumption. However, an alternative possibility is that conceptual structure can in principle drive linear-ordering preferences by itself. This in principle predicts that ordering decisions should be made on an item-by-item basis: a particular object-property pair may have lower PMI than a particular object-numeral pair, and if this indeed impacts how they are represented, then the linguistic tokens should be more likely to be ordered, for example, Adj-Num-N or N-Num-Adj. Note that this completely item-based ordering is not how languages are typically organized. For example, even though numerosity is conceptually more closely related than color to eyes, the ordering observed in English obeys the general category order rather than the item-specific order (e.g. *two blue eyes* as opposed to *blue two eyes*). However, making ordering decisions on the basis of individual items like this may be overly taxing or complex. Category-based generalizations in linguistic representations (like the hierarchy) is a solution to this problem.

Interestingly, however, there is evidence that the categories may actually be represented in a somewhat more fine-grained way. Specifically, it is often suggested that Ad-

jective is not a single category but a set of (hierarchically) related subcategories, for example, quality > size > shape > color > provenance (e.g. Dixon 1982, Cinque 1993). Indeed, while the link between conceptual closeness has not been applied to distinct modifier types, something like it has been suggested as an explanation for patterns of multiple adjective ordering (e.g. Seiler 1978, Bouchard 2002, Champollion 2006), among other factors (see Scontras et al. 2017). For example, Martin (1969) argued that when a phrase contains multiple adjectives, their relative order is influenced by the degree to which they denote properties that are INHERENT OR ESSENTIAL TO THE DENOTED OBJECT. Figure 5b shows average PMI for color/texture adjectives (which tend to pattern the same) and size adjectives in the English corpora.[11] This correctly predicts that color/texture adjectives should generally be ordered closer to the noun than size adjectives are (e.g. *small green vase* is preferred to *green small vase* in English and other languages; Cinque 1993, Scott 2002, Truswell 2009). These more subtle differences in average PMI may also affect word-order preferences, particularly in a task designed to tap into sensitivity to conceptual information, like silent gesture.

5. EXPERIMENT 2. In experiment 2, we test the prediction that adjectives that differ in their average PMI should differ in the extent to which they are ordered homomorphically in spontaneous gesture strings. To evaluate this, we conduct a second silent gesture experiment, directly comparing the improvised gesture strings produced by participants when object properties were higher-PMI textures (*striped* and *spotted*, as in experiment 1) or lower-PMI sizes (*small* and *large*).

5.1. METHOD.

PARTICIPANTS. Participants were forty native English speakers (twenty per condition). The data set from one participant (texture condition) was excluded due to failure to produce gestures containing information for more than one modifier; therefore data from thirty-nine participants were used for analysis.

Participants were randomly assigned to the texture or size condition; they thus gestured only high- or low-PMI adjectives. None had previous knowledge of any sign language.

MATERIALS. The objects featured in experiment 2 were toothbrushes and pencils (see Figure 6). In experiment 1 (where we used squares and triangles) we observed participants using combined gestures (e.g. shape and texture simultaneously). We anticipated that this would be even more common with size (e.g. conveying large and square as a combined gesture). In order to discourage participants from using simultaneous gestures in experiment 2 we therefore used objects—pencils and toothbrushes—that are typically conveyed using associated actions rather than by depicting the shape (Padden et al. 2015). As in experiment 1, objects appeared in groups of four or five. In the high-PMI condition, objects were either striped or spotted (see Fig. 6a). We used texture rather than color because improvising a gesture for a color is relatively difficult compared to a texture like striped or spotted. In the low-PMI (size) condition, objects were either big or small (see Fig. 6b). Location relative to the gesturer was represented by two iPads that displayed the images, one of which was directly in front of the participant, and the other about an arm's length away (see Fig. 2b). The eight different images, presented on two different iPads, together formed sixteen total stimulus items, which were presented twice, in random order for each participant.

[11] For additional information about PMI calculations for specific adjective classes, see appendix §A1.

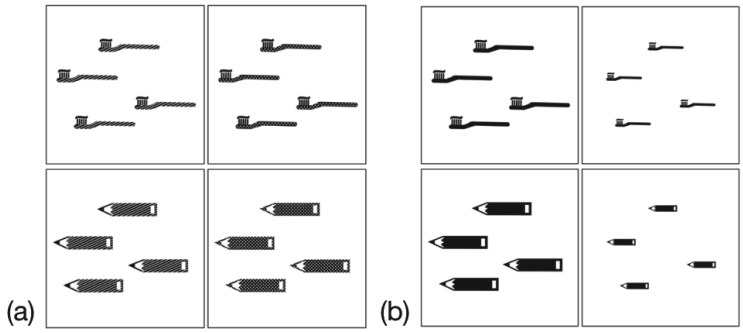

FIGURE 6. Stimulus set for experiment 2: (a) texture condition; (b) size condition.

PROCEDURE. The procedure was identical to experiment 1 with one exception: at the end of experiment 2, we asked participants to reflect on the gesture order they used, and we gave them a single trial to describe using English.

CODING. Example gesture clips for experiment 2 are shown in Figure 7. Gesture strings were coded as for experiment 1. Agreement on gesture order was 90%, on gesture homomorphism 100%. All instances of disagreement were resolved by discussion with a third coder (blind to the hypothesized difference between conditions).

FIGURE 7. Clips from three participants illustrating example gestures. (Top: texture condition, stimulus *four spotted toothbrushes* on proximal iPad, order Dem-N-Adj-Num; middle: texture condition, stimulus *five striped pencils* on distal iPad, order N-Adj-Num-Dem; bottom: size condition, stimulus *four small toothbrushes* on distal iPad, order Dem-N-Num-Adj.)

5.2. RESULTS. As in experiment 1, gesture strings were overwhelmingly homomorphic. Across both conditions 18% were non-homomorphic (1% were ambiguous due to e.g. repetitions). The most commonly used patterns, shown in Figure 8, were Dem-Num-N-A (approximately 400 gestures), Dem-Num-A-N, Dem-N-A-Num, and Dem-N-Num-A (ranging from 150–185 each). This includes the English order (but not the reverse, for example), suggesting some influence of the native language. Participants'

self-reports indicate that few were consciously aware of using English order: very few participants reported that they produced gestures based on how they would have said it in English; most reported having no idea why they used a particular order, or putting the most important or salient information first. The most common orders produced using speech post-experiment were Dem-Num-N-A, Num-A-N-Dem, Dem-Num-A-N, and Num-A-N (with Dem omitted). These overlap partially, but not entirely, with the gesture orders overall, and there is a partial correspondence between the speech order participants used and their gesture order. For example, of the nine participants who produced an order resembling English in speech ((Dem)-Num-A-N), four used this in gesture as well, while the remaining five used a different order (e.g. Dem-Num-N-Num-A, Dem-N-A-Num, Dem-Num-N-A).

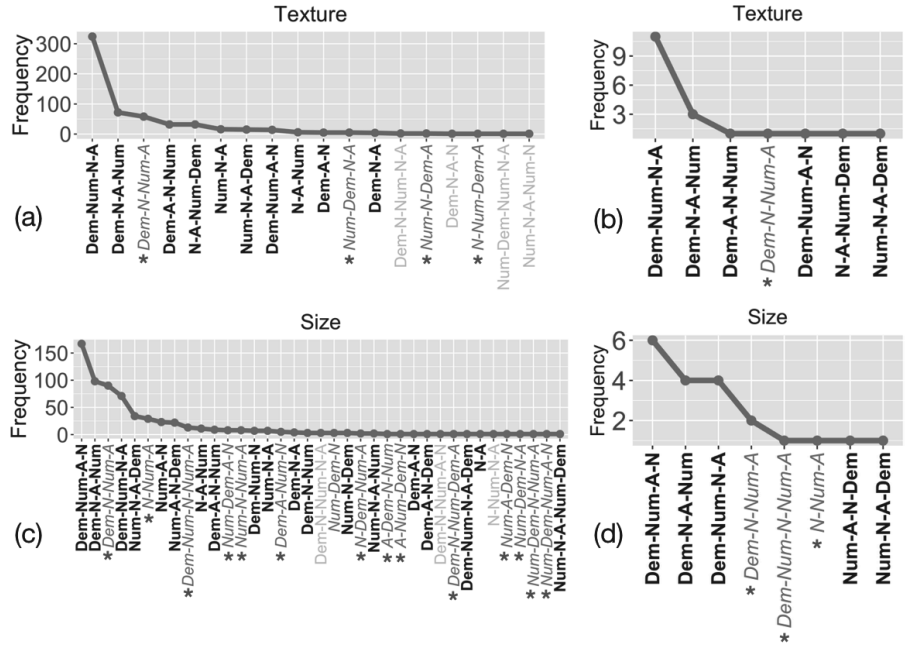

FIGURE 8. Top row: texture condition (high PMI). (a) Raw frequency of each order produced by participants. Bold black patterns are homomorphic; gray italic patterns preceded by stars are non-homomorphic; light gray patterns are ambiguous; elements in curly brackets were gestured simultaneously. (b) Patterns that were used in the majority of at least one participant's gestures (x-axis), and the number of participants using each (y-axis). Bottom row: size condition (low PMI). (c) Raw frequency of each order produced by participants. (d) Patterns used in the majority of at least one participant's gestures (x-axis), and the number of participants using each (y-axis). The overwhelming majority of orders produced were homomorphic, with more non-homomorphic orders in the size condition.

Returning to the question of interest here, Figure 8 suggests that more non-homomorphic orders were produced by participants in the low-PMI size condition. Importantly, the most frequent of these involves placing Num closer to N than A, exactly as predicted. In the texture condition 89% of gestures were homomorphic (11% non); in the size condition 74% were homomorphic (25% non). A logistic mixed-effects model predicting homomorphic order from condition, with participant as a random effect, revealed a significant overall preference for homomorphic order but no difference between conditions, either overall or for A and Num specifically (sum contrast coding, overall scope: inter-

cept $\beta = 3.08 \pm 0.53$, $p < 0.001$, condition $\beta = -0.63 \pm 0.50$, $p = 0.20$; A and Num scope: intercept $\beta = 3.48 \pm 0.64$, $p < 0.001$, condition $\beta = -0.90 \pm 0.57$, $p = 0.11$). Comparing these results to those of experiment 1, there was no difference in use of homomorphic orders between the two texture conditions ($\beta = -1.6 \pm 1.12$, $p = 0.15$), but there was a significant difference between homomorphism in experiment 1 (texture only) and the size condition in experiment 2 ($\beta = -2.89 \pm 1.09$, $p < 0.01$). A comparison between the two texture conditions combined and the size condition in experiment 2 also revealed a significant difference in homomorphism ($\beta = -0.96 \pm 0.44$, $p = 0.03$). In both cases, homomorphic orders were less likely to be used in the size condition. However, the pressure to use a homomorphic order remains strong across conditions.

6. General discussion. Noun phrase word-order patterns documented in the world's languages follow a highly skewed distribution: a small number of patterns are very common, while others are rare or as yet unattested. In previous research, linguists have claimed that this distribution is evidence for the influence of universal organizing principles, at least some of which might be innate (Cinque 2005, Steddy & Samek-Lodovici 2011, Abels & Neeleman 2012, Dryer 2018, Steedman 2018). Here we have explored these universal organizing principles using novel sources of data. First, we found evidence for a preference for homomorphic orders, that is, orders that transparently map between a universal underlying structure and linear order. In line with previous work using artificial (spoken) language experiments (Culbertson & Adger 2014, Martin et al. 2019), we found that gesture strings, improvised in a modality distinct from participants' previous language experience, were overwhelmingly homomorphic.

Second, we used an information-theoretic measure of strength of association computed on natural language corpora (as a proxy for the real world) to show that objects (expressed by Nouns) and their properties (expressed by Adjectives) are more closely related than objects and their numerosities (expressed by Numerals); objects and their numerosities are in turn more closely related than objects and their location or relation to the speaker (expressed by Demonstratives). Specifically, average pointwise mutual information differed among these distinct types, with (Noun, Adjective) pairs consistently highest, and (Noun, Demonstrative) pairs lowest. We argued that conceptual representations reflecting this kind of information could form the basis of a universal linguistic hierarchy (along the lines of Fig. 1b). To confirm the relationship between strength of association and linear order using a more fine-grained distinction, our second gesture experiment manipulated types of properties (conveyed by size and texture adjectives). When the property to be conveyed was on average more similar in terms of PMI to numerosity (size), this resulted qualitatively in more violations of homomorphism. This effect reached significance only when combined with the data from experiment 1, confirming again the strength of the homomorphism bias.

Our findings are therefore consistent with the idea that conceptual structure and linear order are related via homomorphism. However, there remain a number of open questions as to the nature of the link between strength of association and linear order, some of which we have already discussed above. First, in natural language the link between conceptual representations and word order may be mediated by knowledge of (discrete) linguistic categories (like Noun, Adjective, Numeral, and Demonstrative) and by an intermediate level of representation encoding syntactic constituency and semantic composition (see Figure 9). It may be that the bias for homomorphism targets these linguistic representations, rather than the conceptual structure directly. After all, syntactic categories like Noun, Adjective, Numeral, and Demonstrative are linguistic notions,

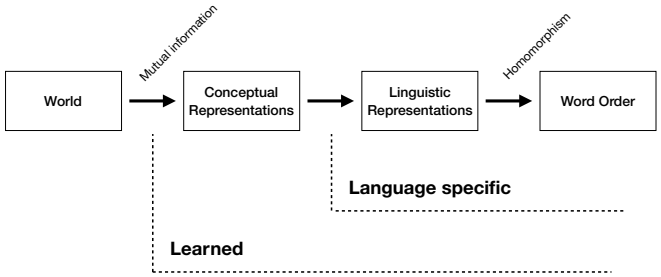

FIGURE 9. How properties of the world shape the word order of languages. Our experimental results and corpus statistics support a view in which conceptual representations are learned by children exposed to statistical properties of the world. This leads to hierarchically structured linguistic representations, which, when linearized homomorphically, predict word-order universals. The difference between conceptual and linguistic representations here relates to categories such as Noun, Adjective, Numeral, and Demonstrative, over which generalizations must be made by the language learner. The 'linguistic representations' box in this figure therefore corresponds to the hierarchy in Fig. 1b. Note that although the homomorphism bias here is operating on representations that are specific to language, we propose that the bias itself arises from domain-general principles of simplicity.

and rules determining linear order in a given language typically appear to target these (rather than individual tokens or pairs of words).

There is, however, another possibility, which does not invoke a homomorphism bias at all. Recent work on word order and language processing suggests that mutual information derived from surface linguistic input (rather than observing the world) may also influence word order directly. Most recently, Hahn et al. (2018) show that mutual information contributes to explaining adjective-ordering preferences gathered from English speakers; when multiple adjectives are present in a phrase, the adjective with higher mutual information tends to be closer to the noun. Following Futrell and Levy (2017), this could be driven by memory-constrained incremental processing. Placing modifier-noun pairs with high mutual information far apart from one another may increase processing effort. Here, high mutual information IN THE CORPUS could in principle be driven by strong conceptual association. This is intriguing work, but deals only with adjective order. Nevertheless, it would be interesting to explore the empirical predictions it would make for the order of other modifiers, and whether it might provide an alternative or complementary approach to noun phrase order grounded primarily in incremental processing considerations rather than a preference for homomorphic mappings to conceptual structure. Importantly, this account would also need to incorporate a notion of category-based generalization that would override the pressure to make ordering decisions on a token-by-token basis. In addition, note that for us, linguistic corpora necessarily reflect the statistical properties of the world (and this is what allows us to use them as a proxy for the purposes of estimating the mutual information between different aspects of the world). As such, we would argue that the ultimate source of the mutual information asymmetries in the surface linguistic input available to children will itself be driven by the same properties of the world that we have appealed to here. In this case, the direction of causation is from mutual information IN THE WORLD to conceptual structure.

To conclude, both the diversity and the similarities among patterns in this simple linguistic domain are critically important. The similarities allow us to relate features of language to general features of cognition, filtered through linguistic representations. We have argued that statistical properties learned from observing the world set up a con-

ceptual asymmetry in which objects are more closely related to their properties than to their numerosities, which are in turn more closely related to objects than their location/relation to the speaker is. This leads to a set of universal underlying hierarchical relations between the linguistic categories expressing these elements: Nouns, Adjectives, Numerals, and Demonstratives. A pressure for transparent mappings between this underlying structure and linear order leads languages to favor a particular set of noun phrase orders—with the adjective closest to, and the demonstrative farthest away from, the noun. There are eight such orders, and they are all among the most robustly attested in the world's languages. This general pressure for homomorphism reflects a type of simplicity, often taken to be a unifying principle of cognitive science (Chater & Vitányi 2003, Culbertson & Kirby 2016). The diversity of patterns actually found illuminates the probabilistic nature of the mechanism linking cognition and linguistic structure, namely cultural evolution.

Appendix

A1. Additional information about PMI calculations. Adjectives, numerals, and demonstratives differ from one another in terms of the size of the linguistic class; there are typically more adjectives than numerals, and a relatively small set of demonstratives. For example, English uses four (encoding distal/proximate and singular/plural distinctions), Latin used six (differentiating 'near to me' and 'near to you' in addition to 'far away'), and other languages like Ilocano have a much larger set (making many distinctions in terms of visibility, continued existence, etc.; Rubino 2000). But most languages tend to use a small number of demonstratives relative to adjectives and numerals. These categories themselves therefore reflect how the world is carved up by the human mind, both in ways that are specific to particular languages and in ways that are likely not. Intuitively, though, even if one were to vastly increase the number of demonstratives—for example, making ever finer distinctions in terms of physical or temporal distance—they will still be relative to the speaker. However, modifier types with fewer category members will typically have lower entropy, and this could in turn lower their PMI since the mutual information of a pair is bounded above by the entropy of the lower-entropy element. Note, however, that for some languages, the set of adjectives and numerals is similar in size, and yet the PMI values still differ (e.g. in our Indonesian corpus, there are 125 unique adjectives and 123 unique numerals, but adjective PMIs are still almost twice as high). Although our hypothesis posits a connection between linear order and PMI (not modifier set size or entropy), we checked whether entropy rather than PMI explains the differences between modifier types in our data. To do this, we randomly sampled small sets of adjectives and numerals from our English corpora and calculated PMIs for each sample (1,000 samples, set size = 4, same as Dem). While entropies go down overall, as expected, the PMI of (Adj, N) pairs is still highest, and the PMI for (Dem, N) the lowest. We also randomly sampled sets holding entropy near constant (a tight range around the entropy for demonstratives in our corpus), and our PMI differences still hold there as well. See Figure A1.

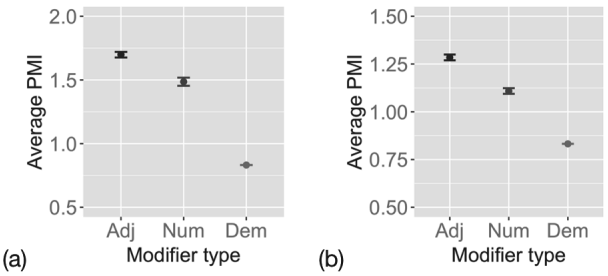

Figure A1. (a) Average PMI values for 1,000 samples of each modifier type, with set size fixed at 4. This means demonstratives are always the same, but adjectives and numerals are randomly sampled from the unique set of each type. (b) Average PMI values for 1,000 entropy-matched samples of each modifier type, set size again fixed to 4.

Average PMI values for color/texture vs. size adjectives were calculated by extracting the PMI values calculated as above for adjectives that matched these types. Adjectives were identified from lists of color, texture (including material), and size adjectives and then supplemented by hand coding. We also computed PMI val-

ues for these adjectives a second way: by recomputing PMI values such that the probability of a given modifier is based on the set of modifiers of that type (i.e. here color or size rather than adjective) modifying a head noun, and the probability of the noun is based on the set of nouns that have that type of modifier. This does not change the result, namely, that PMI values for color/texture are higher than for size.

A2. ADDITIONAL INFORMATION ABOUT CORPORA. All adult-directed corpora are from the Universal Dependencies Treebank 2.0 (Nivre et al. 2018). English child-directed speech data are the dependency-parsed English corpora available on CHILDES (Sagae et al. 2007). We used all languages with available corpora totaling more than 200K words (CHILDES is the largest), plus Basque and Indonesian (approximately 150K each) in order to increase the diversity of the set. All treebank corpora used are listed in Table A1.

LANGUAGE	FAMILY	SOURCE
French	IE, Romance	Sequoia Treebank (Candito & Seddah 2012), French UD Treebank (McDonald et al. 2013), Turin University Parallel Treebank (Sanguinetti & Bosco 2015)
Spanish	IE, Romance	AnCora Treebank (Recasens & Martí 2010), Spanish UD Treebank (McDonald et al. 2013)
Catalan	IE, Romance	AnCora Treebank (Recasens & Martí 2010)
Italian	IE, Romance	Italian UD Treebank (Bosco et al. 2000), Turin University Parallel Treebank (Sanguinetti & Bosco 2015), PoSTWITA-UD (Sanguinetti et al. 2018)
Portuguese	IE, Romance	Portuguese UD Treebank (McDonald et al. 2013), UD Portuguese Treebank (Rademaker et al. 2017)
Romanian	IE, Romance	Romanian UD Treebank (Barbu Mititelu et al. 2016), Romanian Nonstandard UD Treebank (Bobicev et al. 2016)
Latin	IE, Romance	UD Latin PROIEL Treebank (Haug & Jøhndal 2008), Perseus UD Latin Treebank (Bamman & Crane 2011), Index Thomisticus Treebank (Cecchini et al. 2018)
Russian	IE, Slavic	Russian Universal Dependencies Treebank (McDonald et al. 2013), SynTagRus (Dyachenko et al. 2015), UD Russian Taiga (Lyashevkaya et al. 2016)
Czech	IE, Slavic	Czech CAC UD Treebank (Hladká et al. 2008), Czech PDT UD Tree bank (Bejček et al. 2012), Czech CLTT UD Treebank (Kríž et al. 2015), FicTree (Jelínek 2017), Parallel Universal Dependencies Treebank
Polish	IE, Slavic	UD Polish Treebank (Wróblewska & Przepiórkowski 2014), LFG Enhanced UD Treebank of Polish (Patejuk & Przepiórkowski 2018)
Hindi	IE, Indic	Hindi UD Treebank (Palmer et al. 2009)
Ancient Greek	IE, Greek	UD Ancient Greek PROIEL (Haug & Jøhndal 2008), Perseus Universal Dependencies Greek Treebank (Bamman & Crane 2011)
English	IE, Germanic	English Web Treebank (Bies et al. 2012), LinES Parallel Treebank (Ahrenberg 2015), Turin University Parallel Treebank (Sanguinetti & Bosco 2015), Georgetown University Multilayer corpus (Zeldes 2017), Parallel Universal Dependencies Treebank
German	IE, Germanic	German UD Treebank (McDonald et al. 2013)
Swedish	IE, Germanic	Swedish-Talbanken Treebank (Nivre & Bandmann Megyesi 2007), LinES Parallel Treebank (Ahrenberg 2015)
Dutch	IE, Germanic	UD Dutch Alpino Treebank (Van der Beek et al. 2002), UD Lassy Small Treebank (Bouma & Van Noord 2017)
Norwegian	IE, Germanic	LIA Norwegian UD Treebank (Øvrelid & Hohle 2016), Norwegian UD Treebank, Bokmål, Nynorsk (Velldal et al. 2017)
Estonian	Uralic, Finnic	UD Estonian Treebank (Muischnek et al. 2014)
Finnish	Uralic, Finnic	UD FinnTreeBank 1 (Hakulinen et al. 2004), UD Turku Dependency Treebank (Haverinen et al. 2014)
Chinese	Sino-Tibetan	Traditional Chinese UD Treebank (McDonald et al. 2013)
Korean	Korean	Korean UD Treebank (McDonald et al. 2013), KAIST Korean UD Treebank (Chun et al. 2018)
Arabic	Afro-Asiatic, Semitic	NYUAD Arabic UD treebank (Maamouri et al. 2005), Arabic-PADT UD Treebank (Smrž et al. 2008)

(TABLE A1. *Continues*)

LANGUAGE	FAMILY	SOURCE
Japanese	Japanese	Japanese UD Treebank (McDonald et al. 2013)
Indonesian	Austronesian, Malayo-Sumbawan	Indonesian UD Treebank (McDonald et al. 2013)
Basque	Basque	Basque UD Treebank (Aranzabe et al. 2014)

TABLE A1. Corpora used.

Information about word order is not encoded directly in the dependencies, and the overwhelming majority of phrases containing one of these three modifier types in the corpora have only a single modifier (83% in the English treebanks, compared to 14% with two modifiers, and 3% with three or more). Further, almost all phrases (94%) have only a single modifier type. Note that demonstratives are tagged inconsistently across treebank corpora. We extracted them using featural information first, and then manually checked them against grammars for each language in order to assemble the final set. Numeral dependencies are also tagged somewhat inconsistently across corpora (e.g. dates are sometimes tagged as cardinal numerals, some compound number words, such as *two hundred*, are occasionally treated as two distinct words). We programmatically cleaned up the numeral dependencies where this was possible.

REFERENCES

ABELS, KLAUS, and AD NEELEMAN. 2012. Linear asymmetries and the LCA. *Syntax* 15.25–74. DOI: 10.1111/j.1467-9612.2011.00163.x.

ADGER, DAVID. 2003. *Core syntax*. Oxford: Oxford University Press.

AHRENBERG, LARS. 2015. Converting an English-Swedish parallel treebank to universal dependencies. *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, 10–19. Online: https://www.aclweb.org/anthology/W15-2103.

ALEXIADOU, ARTEMIS; LILIANE HAEGEMAN; and MELITA STAVROU. 2007. *Noun phrase in the generative perspective*. Berlin: Mouton de Gruyter. DOI: 10.1515/9783110207491.

ARANZABE, MARIA JESUS; AITZIBER ATUTXA; KEPA BENGOETXEA; ARANTZA DIAZ; IAKES GOENAGA DE ILARRAZA; KOLDO GOJENOLA; and LARRAITZ URIA. 2014. Automatic conversion of the Basque Dependency Treebank to universal dependencies. *Proceedings of the 14th Workshop on Treebanks and Linguistic Theories (TLT14)*, 233–41. Online: http://tlt14.ipipan.waw.pl/index.php/download_file/view/17/152/.

BAMMAN, DAVID, and GREGORY CRANE. 2011. The Ancient Greek and Latin Dependency Treebanks. *Language technology for cultural heritage*, ed. by Caroline Sporleder, Antal van den Bosch, and Kalliopi Zervanou, 79–98. Berlin: Springer. DOI: 10.1007/978-3-642-20227-8_5.

BARBU MITITELU, VERGINICA; RADU ION; RADU SIMIONESCU; ELENA IRIMIA; and CENEL-AUGUSTO PEREZ. 2016. The Romanian treebank annotated according to universal dependencies. *Proceedings of the Tenth International Conference on Natural Language Processing (HrTAL2016)*.

BATES, DOUGLAS M. 2010. lme4: Mixed-effects modeling with R. Madison: University of Wisconsin, MS. Online: http://lme4.r-forge.r-project.org/book.

BEJČEK, EDUARD; JARMILA PANEVOVÁ; JAN POPELKA; PAVEL STRAŇÁK; MAGDA ŠEVČÍKOVÍ; JAN ŠTĚPÁNEK; and ZDENĚK ŽABOKRTSKY. 2012. Prague Dependency Treebank 2.5—A revisited version of PDT 2.0. *Proceedings of COLING 2012*, 231–46. Online: https://www.aclweb.org/anthology/C12-1015.

BIES, ANN; JUSTIN MOTT; COLIN WARNER; and SETH KULICK. 2012. English Web Treebank. LDC2012T13. Philadelphia: Linguistic Data Consortium.

BOBICEV, VICTORI; TUDOR BUMBU; VICTORIA LAZU; VICTORIA MAXIM; and DANIELA ISTRATI. 2016. Folk poetry for computers: Moldovan Codri's ballads parsing. *Proceedings of the 12th international conference 'Linguistic Resources and Tools for Processing the Romanian Language'*, 39–50.

BOSCO, CRISTINA; VINCENZO LOMBARDO; LEONARDO LESMO; and VASSALLO DANIELA. 2000. Building a treebank for Italian: A data-driven annotation schema. *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC 2000)*, 99–105. Online: http://www.lrec-conf.org/proceedings/lrec2000/pdf/220.pdf.

BOUCHARD, DENIS. 2002. *Adjectives, number and interfaces: Why languages vary*. Amsterdam: Elsevier.

Bouma, Gosse, and Gertjan Van Noord. 2017. Increasing return on annotation investment: The automatic construction of a universal dependency treebank for Dutch. *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, 19–26. Online: https://www.aclweb.org/anthology/W17-0403.

Candito, Marie, and Djamé Seddah. 2012. Le corpus Sequoia : Annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical [The Sequoia corpus: Syntactic annotation and use for a parser lexical domain adaptation method]. *Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, vol. 2: TALN*, 321–34. Online: https://www.aclweb.org/anthology/F12-2024.

Cecchini, Flavio Massimiliano; Marco Passarotti; Paola Marongiu; and Daniel Zeman. 2018. Challenges in converting the *Index Thomisticus* treebank into universal dependencies. *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, 27–36. DOI: 10.18653/v1/W18-6004.

Champollion, Lucas. 2006. A game-theoretic account of adjective ordering restrictions. Philadelphia: University of Pennsylvania, ms. Online: https://www.ling.upenn.edu/~champoll/adjective-ordering.pdf.

Chater, Nick, and Paul Vitányi. 2003. Simplicity: A unifying principle in cognitive science? *Trends in Cognitive Sciences* 7.19–22. DOI: 10.1016/S1364-6613(02)00005-0.

Chun, Jayeol; Na-Rae Han; Jena D. Hwang; and Jinho D. Choi. 2018. Building universal dependency treebanks in Korean. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2194–2202. Online: http://www.lrec-conf.org/proceedings/lrec2018/pdf/378.pdf.

Cinque, Guglielmo. 1993. On the evidence for partial N-movement in the Romance DP. *University of Venice Working Papers in Linguistics* 3.21–40.

Cinque, Guglielmo. 2005. Deriving Greenberg's universal 20 and its exceptions. *Linguistic Inquiry* 36.315–32. DOI: 10.1162/0024389054396917.

Culbertson, Jennifer, and David Adger. 2014. Language learners privilege structured meaning over surface frequency. *Proceedings of the National Academy of Sciences* 111.5842–47. DOI: 10.1073/pnas.1320525111.

Culbertson, Jennifer, and Simon Kirby. 2016. Simplicity and specificity in language: Domain-general biases have domain-specific effects. *Frontiers in Psychology* 6:1964. DOI: 10.3389/fpsyg.2015.01964.

Culbertson, Jennifer, and Elissa L. Newport. 2015. Harmonic biases in child learners: In support of language universals. *Cognition* 139.71–82. DOI: 10.1016/j.cognition.2015.02.007.

Culbertson, Jennifer; Paul Smolensky; and Géraldine Legendre. 2012. Learning biases predict a word order universal. *Cognition* 122.306–29. DOI: 10.1016/j.cognition.2011.10.017.

Culbertson, Jennifer; Paul Smolensky; and Colin Wilson. 2013. Cognitive biases, linguistic universals, and constraint-based grammar learning. *Topics in Cognitive Science* 5.392–424. DOI: 10.1111/tops.12027.

Cysouw, Michael. 2010. Dealing with diversity: Towards an explanation of NP-internal word order frequencies. *Linguistic Typology* 14.253–87. DOI: 10.1515/lity.2010.010.

Dixon, Robert M. W. 1982. *Where have all the adjectives gone? And other essays in semantics and syntax*. Berlin: De Gruyter. DOI: 10.1515/9783110822939.

Dryer, Matthew. 2018. On the order of demonstrative, numeral, adjective, and noun. *Language* 94.798–833. DOI: 10.1353/lan.2018.0054.

Dyachenko, P.; L. Iomdin; A. Lazursky; L. Mityushin; O. Podlesskaya; S. Sizov; T. Frolova; and L. Tsinman. 2015. The current state of the deeply annotated corpus of the texts of the Russian language (SynTagRus). *Proceedings of the Russian Language Institute*, 272–300.

Elbourne, Paul. 2008. Demonstratives as individual concepts. *Linguistics and Philosophy* 31.409–66. DOI: 10.1007/s10988-008-9043-0.

Futrell, Richard; Tina Hickey; Aldrin Lee; Eunice Lim; Elena Luchkina; and Edward Gibson. 2015. Cross-linguistic gestures reflect typological universals: A subject-initial, verb-final bias in speakers of diverse languages. *Cognition* 136.215–21. DOI: 10.1016/j.cognition.2014.11.022.

Futrell, Richard, and Roger Levy. 2017. Noisy-context surprisal as a human sentence processing cost model. *Proceedings of the 15th conference of the European Chapter of*

the Association for Computational Linguistics, vol. 1: Long papers, 688–98. Online: https://www.aclweb.org/anthology/E17-1065.

Gibson, Edward; Steven T. Piantadosi; Kimberly Brink; Leon Bergen; Eunice Lim; and Rebecca Saxe. 2013. A noisy-channel account of crosslinguistic word-order variation. *Psychological Science* 24.1079–88. DOI: 10.1177/0956797612463705.

Goldin-Meadow, Susan; Wing Chee So; Asli Özyürek; and Carolyn Mylander. 2008. The natural order of events: How speakers of different languages represent events nonverbally. *Proceedings of the National Academy of Sciences* 105.9163–68. DOI: 10.1073/pnas.0710060105.

Hahn, Michael; Judith Degen; Noah D. Goodman; Dan Jurafsky; and Richard Futrell. 2018. An information-theoretic explanation of adjective ordering preferences. *Proceedings of the 40th annual meeting of the Cognitive Science Society (CogSci 2018)*, 1766–71. Online: https://cogsci.mindmodeling.org/2018/papers/0339/0339.pdf.

Hakulinen, Auli; Maria Vilkuna; Riitta Korhonen; Vesa Koivisto; Riitta Heinonen Tarja; and Irja Alho. 2004. *Iso suomen kielioppi*. Helsinki: Suomalaisen Kirjallisuuden Seura.

Hall, Matthew L.; Rachel I. Mayberry; and Victor S. Ferreira. 2013. Cognitive constraints on constituent order: Evidence from elicited pantomime. *Cognition* 129.1–17. DOI: 10.1016/j.cognition.2013.05.004.

Haug, Dag T. T., and Marius Jøhndal. 2008. Creating a parallel treebank of the old Indo-European Bible translations. *Proceedings of the second workshop on Language Technology for Cultural Heritage Data (LaTeCH 2008)*, 27–34.

Haverinen, Katri; Jenna Nyblom; Timo Viljanen; Veronika Laippala; Samuel Kohonen; Anna Missilä; Stina Ojala; Tapio Salakoski; and Filip Ginter. 2014. Building the essential resources for Finnish: The Turku Dependency Treebank. *Language Resources and Evaluation* 48.493–531. DOI: 10.1007/s10579-013-9244-1.

Heim, Irene, and Angelika Kratzer. 1998. *Semantics in generative grammar*. Oxford: Blackwell.

Hladká, Barbora; Jan Hajič; Jirka Hana; Jaroslava Hlaváčová; Jiří Mírovsky; and Jan Raab. 2008. The Czech Academic Corpus 2.0 guide. *The Prague Bulletin of Mathematical Linguistics* 89.41–96.

Hurford, James R. 1987. *Language and number: The emergence of a cognitive system*. Oxford: Blackwell.

Jelínek, Tomáš. 2017. FicTree: A manually annotated treebank of Czech fiction. *ITAT 2017 Proceedings*, 181–85. Online: http://ceur-ws.org/Vol-1885/181.pdf.

Jurafsky, Daniel, and James H. Martin. 2019. *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. 3rd edn. draft. Online: https://web.stanford.edu/~jurafsky/slp3/.

Kríž, Vincent; Barbora Hladká; and Zdeňka Urešová. 2015. Czech Legal Text Treebank. Prague: LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University. Online: http://hdl.handle.net/11234/1-1516.

Kucker, Sarah C.; Larissa K. Samuelson; Lynn K. Perry; Hanako Yoshida; Eliana Colunga; Megan G. Lorenz; and Linda B. Smith. 2019. Reproducibility and a unifying explanation: Lessons from the shape bias. *Infant Behavior and Development* 54.156–65. DOI: 10.1016/j.infbeh.2018.09.011.

Landau, Barbara; Linda B. Smith; and Susan S. Jones. 1988. The importance of shape in early lexical learning. *Cognitive Development* 3.299–321. DOI: 10.1016/0885-2014(88)90014-7.

Lyashevkaya, Olga; Kira Droganova; Daniel Zeman; Maria Alexeeva; Tatiana Gavrilova; Nina Mustafina; and Elena Shakurova. 2016. Universal dependencies for Russian: A new syntactic dependencies tagset. Higher School of Economics Research Paper No. WP BRP 44/LNG/2016, Moscow.

Maamouri, Mohamed; Ann Bies; Tim Buckwalter; Hubert Jin; and Wigdan Mekki. 2005. Arabic Treebank: Part 3 (full corpus) v 2.0 (MPG + syntactic analysis). LDC2005T20. Philadelphia: Linguistic Data Consortium.

Martin, Alexander; A. Holtz; Klaus Abels; David Adger; and Jennifer Culbertson. 2020. Experiment evidence for the influence of structure and meaning on linear order in the noun phrase. *Glossa: A Journal of General Linguistics*, to appear.

Martin, Alexander; Theeraporn Ratitamkul; Klaus Abels; David Adger; and Jennifer Culbertson. 2019. Cross-linguistic evidence for cognitive universals in the noun phrase. *Linguistics Vanguard* 5(1):20180072. DOI: 10.1515/lingvan-2018-0072.

Martin, James E. 1969. Semantic determinants of preferred adjective order. *Journal of Verbal Learning and Verbal Behavior* 8.697–704. DOI: 10.1016/S0022-5371(69)80032-0.

McDonald, Ryan; Joakim Nivre; Yvonne Quirmbach-Brundage; Yoav Goldberg; Dipanjan Das; Kuzman Ganchev; Keith Hall; Slav Petrov; Hao Zhang; Oscar Täckström; et al. 2013. Universal dependency annotation for multilingual parsing. *Proceedings of the 51st annual meeting of the Association for Computational Linguistics, vol. 2: Short papers*, 92–97. Online: https://www.aclweb.org/anthology/P13-2017.

Muischnek, Kadri; Kaili Müürisep; Tiina Puolakainen; Eleri Aedmaa; Riin Kirt; and Dage Särg. 2014. Estonian Dependency Treebank and its annotation scheme. *Proceedings of 13th Workshop on Treebanks and Linguistic Theories (TLT13)*, 285–91. Online: http://tlt13.sfs.uni-tuebingen.de/tlt13-proceedings.pdf.

Nivre, Joakim; Mitchell Abrams; Željko Agić; Lars Ahrenberg; Lene Antonsen; Maria Jesus Aranzabe; Gashaw Arutie; Masayuki Asahara; Luma Ateyah; Mohammed Attia; et al. 2018. Universal Dependencies 2.2. Prague: LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. Online: http://hdl.handle.net/11234/1-2837.

Nivre, Joakim, and Beáta Bandmann Megyesi. 2007. Bootstrapping a Swedish treebank using cross-corpus harmonization and annotation projection. *Proceedings of the 6th Workshop on Treebanks and Linguistic Theories (TLT6)*, 97–102. Online: http://tlt07.uib.no/papers/11.pdf.

Øvrelid, Lilja, and Petter Hohle. 2016. Universal dependencies for Norwegian. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, 1579–85. Online: http://www.lrec-conf.org/proceedings/lrec2016/pdf/462_Paper.pdf.

Padden, Carol; So-One Hwang; Ryan Lepic; and Sharon Seegers. 2015. Tools for language: Patterned iconicity in sign language nouns and verbs. *Topics in Cognitive Science* 7.81–94. DOI: 10.1111/tops.12121.

Palmer, Martha; Rajesh Bhatt; Bhuvana Narasimhan; Owen Rambow; Dipti Misra Sharma; and Fei Xia. 2009. Hindi syntax: Annotating dependency, lexical predicate-argument structure, and phrase structure. *Proceedings of ICON-2009: 7th International Conference on Natural Language Processing*, 14–17.

Partee, Barbara H. 1987. Noun phrase interpretation and type-shifting principles. *Studies in discourse representation theory and the theory of generalized quantifiers*, ed. by Jeroen Groenendijk, Dick de Jongh, and Martin Stokhof, 115–43. Dordrecht: Foris.

Partee, Barbara H. 1988. Many quantifiers. *Eastern States Conference on Linguistics (ESCOL)* 5.383–402.

Patejuk, Agnieszka, and Adam Przepiórkowski. 2018. *From lexical functional grammar to enhanced universal dependencies: Linguistically informed treebanks of Polish*. Warsaw: Institute of Computer Science, Polish Academy of Sciences. Online: http://nlp.ipipan.waw.pl/Bib/pat:prz:18:book.pdf.

Piantadosi, Steven T., and Edward Gibson. 2014. Quantitative standards for absolute linguistic universals. *Cognitive Science* 38.736–56. DOI: 10.1111/cogs.12088.

R Core Team. 2017. R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. Online: https://www.r-project.org/.

Rademaker, Alexandre; Fabricio Chalub; Livy Real; Cláudia Freitas; Eckhard Bick; and Valeria de Paiva. 2017. Universal dependencies for Portuguese. *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, 197–206. Online: http://aclweb.org/anthology/W17-6523.

Recasens, Marta, and M. Antònia Martí. 2010. AnCora-CO: Coreferentially annotated corpora for Spanish and Catalan. *Language Resources and Evaluation* 44.315–45. DOI: 10.1007/s10579-009-9108-x.

Rijkhoff, Jan. 1990. Explaining word order in the noun phrase. *Linguistics* 28.5–42. DOI: 10.1515/ling.1990.28.1.5.

Rijkhoff, Jan. 2004. *The noun phrase*. Oxford: Oxford University Press.

Rubino, Carl Ralph Galvez. 2000. *Ilocano dictionary and grammar: Ilocano–English, English–Ilocano*. Honolulu: University of Hawai'i Press.
Sagae, Kenji; Eric Davis; Alon Lavie; Brian MacWhinney; and Shuly Wintner. 2007. High-accuracy annotation and parsing of CHILDES transcripts. *Proceedings of the Workshop on Cognitive Aspects of Computational Language Acquisition*, 25–32. Online: https://www.aclweb.org/anthology/W07-0604.
Sanguinetti, Manuela, and Cristina Bosco. 2015. PartTUT: The Turin University parallel treebank. *Harmonization and development of resources and tools for Italian natural language processing within the PARLI project*, ed. by Roberto Basili, Cristina Bosco, Rodolfo Delmonte, Alessandro Moschitti, and Maria Simi, 51–69. Cham: Springer. DOI: 10.1007/978-3-319-14206-7_3.
Sanguinetti, Manuela; Cristina Bosco; Lavelli Alberto; Alessandro Mazzei; and Tamburini Fabio. 2018. PoSTWITA-UD: An Italian Twitter treebank in universal dependencies. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 1768–75. Online: http://www.lrec-conf.org/proceedings/lrec2018/pdf/636.pdf.
Schouwstra, Marieke, and Henriëtte de Swart. 2014. The semantic origins of word order. *Cognition* 131.431–36. DOI: 10.1016/j.cognition.2014.03.004.
Schouwstra, Marieke; Kenneth Smith; and Simon Kirby. 2016. From natural order to convention in silent gesture. *The evolution of language: Proceedings of the 11th International Conference (EVOLANG11)*, 525–26. Online: http://evolang.org/neworleans/papers/67.html.
Scontras, Gregory; Judith Degen; and Noah D. Goodman. 2017. Subjectivity predicts adjective ordering preferences. *Open Mind: Discoveries in Cognitive Science* 1.53–66. DOI: 10.1162/OPMI_a_00005.
Scott, Gary-John. 2002. Stacked adjectival modification and the structure of nominal phrases. *Functional structure in DP and IP: The cartography of syntactic structures, vol. 1*, ed. by Guglielmo Cinque, 91–120. Oxford: Oxford University Press.
Seiler, Hansjakob. 1978. Determination: A functional dimension for interlanguage comparison. *Language universals: Papers from the conference at Gummersbach/Cologne, Germany, October 3–8, 1976*, ed. by Hansjakob Seiler, 301–28. Tübingen: Gunter Narr.
Smrž, Otakar; Viktor Bielický; Iveta Kouřilová; Jakub Kráčmar; Jan Hajič; and Petr Zemánek. 2008. Prague Arabic Dependency Treebank: A word on the million words. *Proceedings of the Workshop on Arabic and Local Languages (LREC 2008)*, 16–23.
Steddy, Sam, and Vieri Samek-Lodovici. 2011. On the ungrammaticality of remnant movement in the derivation of Greenberg's universal 20. *Linguistic Inquiry* 42.445–69. DOI: 10.1162/LING_a_00053.
Steedman, Mark. 2018. A formal universal of natural language grammar. Edinburgh: University of Edinburgh, ms.
Truswell, Robert. 2009. Attributive adjectives and nominal templates. *Linguistic Inquiry* 40.525–33. DOI: 10.1162/ling.2009.40.3.525.
Van der Beek, Leonoor; Gosse Bouma; Rob Malouf; and Gertjan Van Noord. 2002. The Alpino Dependency Treebank. *Language and Computers* 45.8–22.
Velldal, Erik; Lilja Øvrelid; and Petter Hohle. 2017. Joint UD parsing of Norwegian Bokmål and Nynorsk. *Proceedings of the 21st Nordic Conference on Computational Linguistics*, 1–10. Online: https://www.aclweb.org/anthology/W17-0201/.
Wróblewska, Alina, and Adam Przepiórkowski. 2014. Projection-based annotation of a Polish dependency treebank. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, 2306–12. Online: http://www.lrec-conf.org/proceedings/lrec2014/pdf/538_Paper.pdf.
Zeldes, Amir. 2017. The GUM corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation* 51.581–612. DOI: 10.1007/s10579-016-9343-x.

Centre for Language Evolution
University of Edinburgh
3 Charles St.
Edinburgh EH8 9AD, UK
[jennifer.culbertson@ed.ac.uk]
[Marieke.Schouwstra@ed.ac.uk]
[simon.kirby@ed.ac.uk]

[Received 29 August 2019;
revision invited 30 December 2019;
revision received 10 March 2020;
accepted pending revisions 29 March 2020;
revision received 31 March 2020;
accepted 31 March 2020]